SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models

Year: 2023 | Citations: 631 | Authors: Potsawee Manakul, Adian Liusie, M. Gales

---

Abstract

Generative Large Language Models (LLMs) such as GPT-3 are capable of generating highly fluent responses to a wide variety of user prompts. However, LLMs are known to hallucinate facts and make non-factual statements which can undermine trust in their output. Existing fact-checking approaches either require access to the output probability distribution (which may not be available for systems such as ChatGPT) or external databases that are interfaced via separate, often complex, modules. In this work, we propose"SelfCheckGPT", a simple sampling-based approach that can be used to fact-check the responses of black-box models in a zero-resource fashion, i.e. without an external database. SelfCheckGPT leverages the simple idea that if an LLM has knowledge of a given concept, sampled responses are likely to be similar and contain consistent facts. However, for hallucinated facts, stochastically sampled responses are likely to diverge and contradict one another. We investigate this approach by using GPT-3 to generate passages about individuals from the WikiBio dataset, and manually annotate the factuality of the generated passages. We demonstrate that SelfCheckGPT can: i) detect non-factual and factual sentences; and ii) rank passages in terms of factuality. We compare our approach to several baselines and show that our approach has considerably higher AUC-PR scores in sentence-level hallucination detection and higher correlation scores in passage-level factuality assessment compared to grey-box methods.